WARWICK BUSINESS SCHOOL
THE UNIVERSITY OF WARWICK

Masters Programmes: Group Assignment Cover Sheet

Student Numbers: Please list numbers of all group members	**5619764 5669647 5623415 5668506 5673603**
Module Code:	**IB9BW0**
Module Title:	**Analytics in Practice**
Submission Deadline:	**02 Dec 2024, 12:00**
Date Submitted:	**Sunday 1st December**
Word Count:	**1,930 words**
Number of Pages:	**13**
Question Attempted: *(question number/title, or description of assignment)*	**Predict positive customer reviews for eCommerce Platform**
Have you used Artificial Intelligence (AI) in any part of this assignment?	**Yes**

Academic Integrity Declaration

We're part of an academic community at Warwick. Whether studying, teaching, or researching, we're all taking part in an expert conversation which must meet standards of academic integrity. When we all meet these standards, we can take pride in our own academic achievements, as individuals and as an academic community.

Academic integrity means committing to honesty in academic work, giving credit where we've used others' ideas and being proud of our own achievements.

In submitting my work, I confirm that:

- I have read the guidance on academic integrity provided in the Student Handbook and understand the University regulations in relation to Academic Integrity. I am aware of the potential consequences of Academic Misconduct.
- I declare that this work is being submitted on behalf of my group and is all our own, , except where I have stated otherwise.
- No substantial part(s) of the work submitted here has also been submitted by me in other credit bearing assessments courses of study (other than in certain cases of a resubmission of a piece of work), and I acknowledge that if this has been done this may lead to an appropriate sanction.
- Where a generative Artificial Intelligence such as ChatGPT has been used I confirm I have abided by both the University guidance and specific requirements as set out in the Student Handbook and the Assessment brief. I have clearly acknowledged the use of any generative Artificial Intelligence in my submission, my reasoning for using it and which generative AI (or AIs) I have used. Except where indicated the work is otherwise entirely my own.
- I understand that should this piece of work raise concerns requiring investigation in relation to any of points above, it is possible that other work I have submitted for assessment will be checked, even if marks (provisional or confirmed) have been published.
- Where a proof-reader, paid or unpaid was used, I confirm that the proof-reader was made aware of and has complied with the University's proofreading policy.

Upon electronic submission of your assessment you will be required to agree to the statements above

Table of Contents

Masters Programmes: Group Assignment Cover Sheet 1

1. Introduction 3

2. Data Exploration, Cleaning, and Feature Engineering 3

2.1 Data Description: 3

2.2 Exploratory Data Analysis (EDA) 4

2.3 Data Cleaning 6

2.4 Feature Engineering 6

3. Model Selection and Evaluation 6

3.1 Logistic Regression 7

3.2 Support Vector Machine (SVM) 7

3.3 Ensemble Methods 7

3.3.1 Random Forest Classification (RFC) 7

3.3.2 Gradient Boosting (GBDT) 7

3.3.3 Extreme Gradient Boost (XGBoost) 8

4. Model evaluation 8

5. Conclusion 10

6. Recommendation: 10

7. Challenge 11

7.1 Language Barrier in Data Cleaning 11

7.2 Addressing Bias in Feature Selection 11

7.3 Balancing Recall Scores for Positive and Negative Reviews 11

Reference List 12

1. Introduction

Positive reviews are essential for the success of any business, as they build trust, influence purchasing decisions, enhance customer loyalty, and ultimately drive revenue growth. However, maintaining a steady stream of positive feedback is a challenging endeavour. For Nile, a leading eCommerce platform in South America, consistently receiving positive reviews is crucial for maintaining a competitive edge in a dynamic market. This report focuses on developing a robust predictive model to accurately identify customers likely to provide positive feedback, enabling Nile to optimise its resources and strengthen its competitive position. The following sections outline the step-by-step process of building the predictive machine learning model, evaluating its performance using key metrics, and offering actionable recommendations to integrate the model into Nile's business operations, ensuring measurable business impact and strategic growth.

2. Data Exploration, Cleaning, and Feature Engineering

2.1 Data Description:

Nile provided a total of eight tables from their database, covering key aspects such as customers, sellers, products, and reviews. The database schema (Figure 2.1) illustrates how these tables are interconnected.

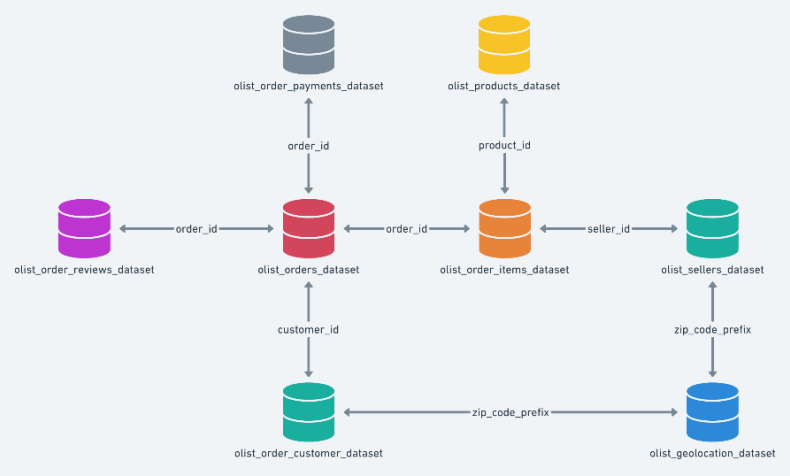

Figure 2.1: Nile's Database Schema

Selected columns (Table 2.1) which are vital for analysing customers' reviews and unique keys (fields ending with "_id") are crucial for joining tables and serving as input features for machine learning. After identifying the necessary columns, 6 tables were merged using left joins on their unique keys.

Table 2.1: Selected Columns of Each Database

Table Name	Selected Columns	Description
olist_order_review_dataset	review score	A vital variable for developing the predictive model.
olist_orders_dataset	order status, order approved at, order estimated delivery date, order delivered customer date	Order progress, approval time, expected delivery date, and actual delivery date.
olist_order_customer_dataset	customer unique id, customer state	Identification of unique customers and their respective locations.
olist_order_payment_dataset	payment type	Details the payment methods used.
olist_order_items_dataset	order item id, price	Sequence of items purchased within a single order and product prices.
olist_products_dataset	product category name, product name length, product description length, product photos quantity	Product-specific details, such as category, name length, description length, and photo count.
olist_sellers_dataset	No essential column	Not utilised in this analysis.
olist_geolocation_dataset	No essential column	

2.2 Exploratory Data Analysis (EDA)

Given that review scores as the target variable for machine learning, a pie chart (Figure 2.2) was created to examine the distribution of positive versus negative reviews, revealing a ratio of approximately 3:1.

Upon reviewing the dataset, orders with statuses other than 'delivered' contain substantial missing information. Additionally, a set of bar charts (Figure 2.3) highlights a significant proportion of negative reviews associated with undelivered orders. These insights indicate that focusing on delivered orders is more relevant for analysing Nile's customer reviews. Consequently, only delivered orders were retained in the combined dataset.

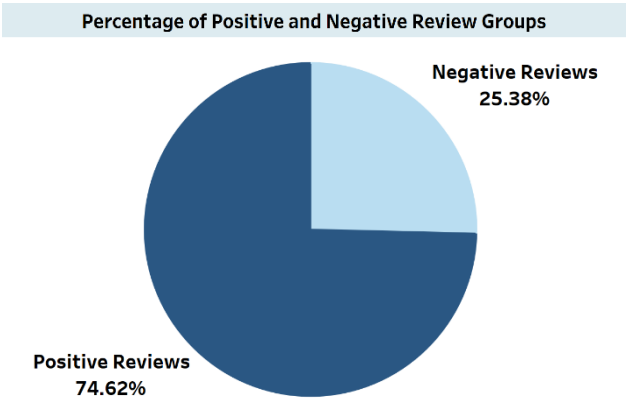

Figure 2.2: Percentage of Positive and Negative Reviews

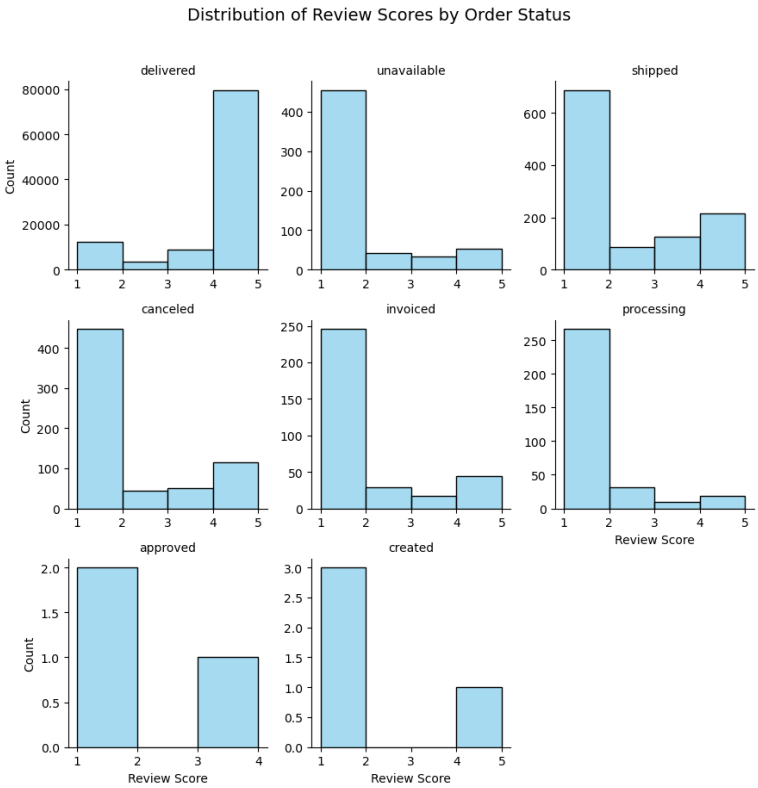

Figure 2.3: Distribution of Review Scores by Order Status

2.3 Data Cleaning

To ensure data integrity, rows with missing values were removed, and duplicates were dropped to eliminate redundant entries, such as identical reviews or multiple purchases of the same product within a single order. This step prevents overrepresentation of certain review scores and preserves the overall distribution of review ratings.

2.4 Feature Engineering

Feature engineering focused on enhancing the dataset to improve model performance. First of all, the definition of negative reviews (Class 0) is review score 1 to 3 and positive reviews (Class 1) are review score 4 to 5. Then, categorical data were converted into numeric representations using dummy variables for columns starting with "type_", "state_", and "category_".

Additional features were derived based on domain-specific insights:

- Product categories were grouped into 10 broader categories to reduce complexity.
- Delivery-related features were added, including the number of days between order approval and delivery, and whether the delivery was earlier than estimated.
- Aggregated features, such as the total price of each order and customer purchase frequency, were also included.

To standardise the dataset, a min-max scaler is applied, ensuring all features fit within a uniform range suitable for machine learning models. Finally, the dataset is split into training and testing subsets with an 80:20 ratio to enable robust evaluation of model performance in subsequent sections.

3. Model Selection and Evaluation

As previously mentioned, this predictive model aims to determine whether future customers will leave positive or negative reviews, representing a binary classification problem. Several methods have proven promising performance in encountering a binary classification problem, including logistic regression, support vector machine (SVM), random forest, and ensemble methods (which are gradient boosting and extreme gradient boost).

3.1 Logistic Regression

Logistic regression is a statistical method commonly used for binary classification tasks. The linearity between dependent and independent values can significantly affect the logistic regression's performance (Nusinovici et al., 2020). Because of the data's nonlinearity, poor performance and accuracy are foreseen during the modelling development.

3.2 Support Vector Machine (SVM)

Because of the characteristics of the nonlinear problem, SVM, a kernel technique mapping the data into a higher-dimensional space (Awad and Khanna, 2015), may be one of the selection models. However, the 3:1 ratio of positive to negative reviews, which indicates an imbalance in the data, is experimented with by current studies that severely hampered results (Cervantes et al., 2020).

3.3 Ensemble Methods

3.3.1 Random Forest Classification (RFC)

RFC executes efficiently on an enormous number of unequal databases and proficiently handles missing data (More and Rana, 2017).

Table 3.1: Evaluation metrics of RFC

	Accuracy	**Precision**	**Recall**	**F1 score**
RFC – Macro	0.81	0.78	0.60	0.62
RFC – Class 1		0.81	0.98	0.89
RFC – Class 0		0.76	0.23	0.34

3.3.2 Gradient Boosting (GBDT)

According to a study by Yap et al.(2013), the most significant feature of conducting boosting is its adaptive nature, allowing it to effectively deal with the nonlinear data structure by adding a set of base learners. Moreover, boosting is also helpful in targeting the imbalance problem in the data, along with providing better performance and efficiency (Burez and Van den Poel, 2009).

Table 3.2: Evaluation metrics of GBDT

	Accuracy	**Precision**	**Recall**	**F1 score**
GBDT – Macro	0.81	0.73	0.62	0.63
GBDT – Class 1		0.82	0.96	0.88
GBDT – Class 0		0.64	0.28	0.39

3.3.3 Extreme Gradient Boost (XGBoost)

Through hundreds of data mining competitions, XGBoost has demonstrated its efficiency, providing a faster and more efficient system. These attributes make XGBoost a suitable choice for expanding the scale of the model deployment, instilling confidence in its performance.

Table 3.3: Evaluation metrics of XGBoost

	Accuracy	**Precision**	**Recall**	**F1 score**
XGBoost – Macro	0.81	0.72	0.61	0.63
XGBoost – Class 1		0.82	0.96	0.88
XGBoost – Class 0		0.61	0.27	0.38

To sum up, due to the characteristics of the nonlinear problem and unbalanced review scores, the ensemble methods, which are slightly affected by bias in minor groups or dealing with enormous noisy data, perfectly fit into the predictive model.

4. Model evaluation

The macro metrics, such as precision, recall, and f1-score, compute the metrics independently for each class and then take the average (Bagui and Li, 2021). Therefore, the weights of the majority and minority are equal and not likely affected by high-probability groups. The comparison among these metrics (Table 4.1) can genuinely reflect an averaged predicting performance instead of reflecting only the majority. Another table (Table 4.2) presents the respective metrics by class 1 and 0. From these two tables, the overall metrics remain consistent across the analysis of all three models, indicating that they already exhibit excellent predictive accuracy, making it difficult to determine a clear winner based on the slight differences in scores.

Table 4.1: Evaluation metrics by machine learning model - Macro

	Accuracy	Precision	Recall	F1 score
RFC	0.81	0.78	0.60	0.62
GBDT	0.81	0.73	0.62	0.64
XGBoost	0.81	0.76	0.62	0.64

Table 4.2: Evaluation metrics by machine learning model and class

	Precision	Recall	F1 score
RFC - Class 1	0.81	0.98	0.89
GBDT – Class 1	0.82	0.96	0.88
XGBoost – Class 1	0.82	0.97	0.89
RFC - Class 0	0.75	0.23	0.35
GBDT – Class 0	0.64	0.28	0.39
XGBoost – Class 0	0.70	0.27	0.39

As in the data exploration section, the pie chart (Figure 2.2) presents an unequal proportion between positive and negative reviews. Therefore, a SMOTE (Synthetic Minority Over-sampling Technique) method is used to sample on minorities. The sampled data will be balanced and improve the performance of XGBoost and GBDT. However, the application of SMOTE did not yield substantial changes in the model's performance metrics, such as accuracy, recall, and F1-score, which remained nearly the same.

The limited impact of SMOTE can be attributed to several potential factors. First, the class imbalance may not be severe enough to require significant changes in the model's behaviour. Research indicates that SMOTE tends to have more impact when the imbalance is more extreme (Pradipta et al., 2021). Moreover, the quality of the synthetic instances generated by SMOTE could be a factor; if these new samples are not representative or realistic, they might not significantly improve the model's performance (He & Garcia, 2009).

Moreover, judging by the feature importance (Figure 4.1), the categorical columns, including payment type, customer states, and ten general product category groups, are minor in the XGBoost model. A study stated that simple models with boosting ensembles perform better than complex ones in most cases (Galar et al., 2012). The categorical columns are excluded to build a good behaviour model.

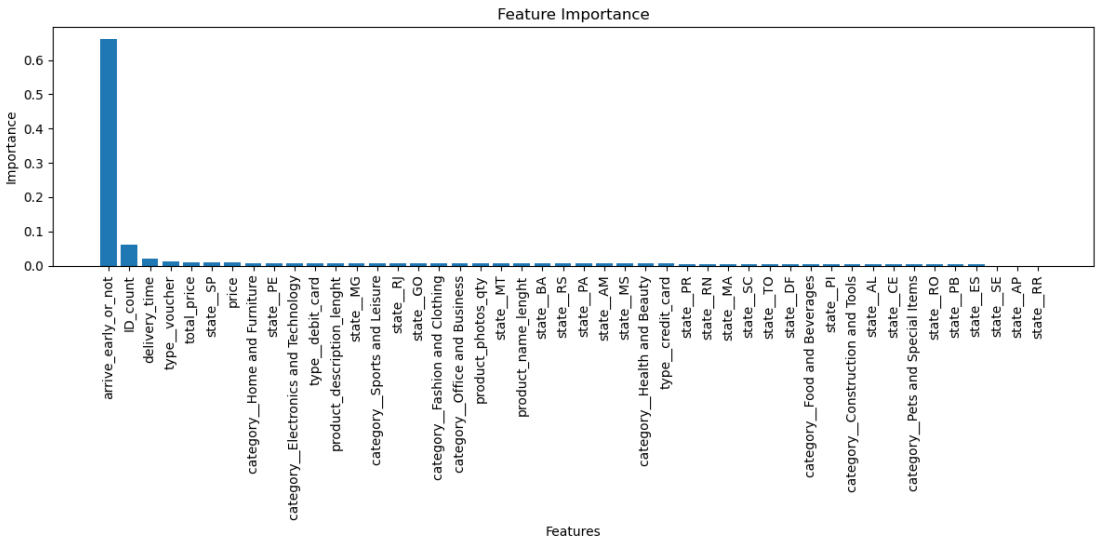

Figure 4.1: XGBoost's Feature Importance

5. Conclusion

This technical report has demonstrated how machine learning can meet Nile's business need of maintaining positive reviews. During the model selection process, we have experimented with several kinds of Machine learning models, including Logistic Regression, Random Forest, Gradient boosting and Extreme Gradient Boosting (XGBoost). After comprehensive evaluation, XGBoost was selected as the preferred algorithm due to its speed and efficiency, as highlighted by Chen & Guestrin (2016), who note its ability to handle large datasets without compromising computational efficiency. Its scalability ensures reliability when Nile's data input increases in the future, making it the optimal solution for predictive modelling and long-term business value.

6. Recommendation:

Based on the results of feature importance analysis (Figure 4.1), it is evident that time-related factors, such as "delivery time" and "whether the package arrives earlier or later than expected" influence the review outcomes significantly. Therefore, it is recommended that Nile prioritize optimizing delivery processes to reduce delivery times and enhance logistical efficiency.

Additionally, launching a next-day delivery service is suggested to Nile as a strategic initiative. This service has the potential to drive short-term sales growth and if executed effectively, could lead to a measurable increase in the amount of positive customer reviews. This recommendation aligns with the observed correlation between prompt delivery and

customer satisfaction, as highlighted by the analysis.

7. Challenge

7.1 Language Barrier in Data Cleaning

Review comments are in Portuguese, but no suitable translation tool is available, complicating sentiment analysis. This hinders accurate alignment between review scores and feedback content, potentially affecting model reliability.

7.2 Addressing Bias in Feature Selection

Feature engineering was based on subjective preferences rather than data-driven methods, risking bias and overfitting. Including irrelevant features may reduce predictive power and hinder generalization(Hastie, Tibshirani, & Friedman, 2009).

7.3 Balancing Recall Scores for Positive and Negative Reviews

The model achieves high recall (0.98) for positive reviews but underperforms (0.21) for negative reviews (Table 7.1), limiting insights into dissatisfied customers. Cost-sensitive learning (Weiss et al., 2007) can improve negative review detection without sacrificing positive review performance, aligning with strategic goals.

Table 7.1: Evaluation metrics by machine learning model and class

	Precision	Recall	F1 score
Random Forest - Class 1	0.81	0.98	0.89
Boosting – Class 1	0.82	0.96	0.88
XGBoost – Class 1	0.82	0.97	0.89
Random Forest - Class 0	0.75	0.23	0.35
Boosting – Class 0	0.64	0.28	0.39
XGBoost – Class 0	0.70	0.27	0.39

Reference List

Awad, M. and Khanna, R. (2015). Support Vector Machines for Classification. *Efficient Learning Machines*, pp.39–66. doi:https://doi.org/10.1007/978-1-4302-5990-9_3.

Bagui, S. and Li, K. (2021). Resampling imbalanced data for network intrusion detection datasets. *Journal of Big Data*, 8(1). doi:https://doi.org/10.1186/s40537-020-00390-x.

Burez, J. and Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3), pp.4626–4636. doi:https://doi.org/10.1016/j.eswa.2008.05.027.

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L. and Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, [online] 408(1), pp.189–215. doi:https://doi.org/10.1016/j.neucom.2019.10.118.

Chen, T. and Guestrin, C. (2016). XGBoost: a Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, pp.785–794. doi:https://doi.org/10.1145/2939672.2939785.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. and Herrera, F. (2012). A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4), pp.463–484. doi:https://doi.org/10.1109/tsmcc.2011.2161285.

Guyon, I. and De, A. (2003). An Introduction to Variable and Feature Selection André Elisseeff. *Journal of Machine Learning Research*, [online] 3, pp.1157–1182. Available at: https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf?trk=public_post_comment-text.

Hastie, T., Tibshirani, R. and Friedman, J. (2017). The elements of statistical learning : data mining, inference, and prediction. *Hoasen.edu.vn*. [online] doi:https://doi.org/978-0-387-84858-7.

He, H. and Garcia, E.A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), pp.1263–1284. doi:https://doi.org/10.1109/tkde.2008.239.

More, A.S. and Rana, D.P. (2017). Review of random forest classification techniques to resolve data imbalance. *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)*. doi:https://doi.org/10.1109/icisim.2017.8122151.

Nusinovici, S., Tham, Y.C., Chak Yan, M.Y., Wei Ting, D.S., Li, J., Sabanayagam, C., Wong, T.Y. and Cheng, C.-Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology*, [online] 122, pp.56–69. doi:https://doi.org/10.1016/j.jclinepi.2020.03.002.

Pradipta, G.A., Wardoyo, R., Musdholifah, A., Sanjaya, I.N.H. and Ismail, M. (2021). *SMOTE for Handling Imbalanced Data Problem : A Review*. [online] IEEE Xplore. doi:https://doi.org/10.1109/ICIC54025.2021.9632912.

Weiss, G., Mccarthy, K. and Zabar, B. (n.d.). *Cost-Sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs?* [online] Available at: https://storm.cis.fordham.edu/~gweiss/papers/dmin07-weiss.pdf [Accessed 1 Dec. 2024].

Yap, B.W., Rani, K.A., Rahman, H.A.A., Fong, S., Khairudin, Z. and Abdullah, N.N. (2013). An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets. *Lecture Notes in Electrical Engineering*, 285, pp.13–22. doi:https://doi.org/10.1007/978-981-4585-18-7_2.